REPORT DOCUMENTATION PAGE

Form Approved OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
(28-04-2008)	Final Report	May 2008 - November 2008

4. TITLE AND SUBTITLE
Enhanced Cutting Plane Techniques for Bi-level Optimization Algorithms

5a. CONTRACT NUMBER

5b. GRANT NUMBER
FA9550-07-1-0404

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
J. Cole Smith

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of Florida
Office of Engineering Research
339 Weil Hall
PO Box 116550
Gainesville, FL 32611-6550

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Air Force Office of Scientific Research
875 North Randolph Street
Arlington, VA 22203

10. SPONSOR/MONITOR'S ACRONYM(S)
AFOSR

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
AFRL-SR-AR-TR-08-0253

12. DISTRIBUTION / AVAILABILITY STATEMENT
Distribution A: unlimited

13. SUPPLEMENTARY NOTES
N/A

14. ABSTRACT
Bi-level optimization algorithms have been developed for use in a variety of problems, including network design and defense systems and optimization under uncertainty, and network interdiction applications. These algorithms often employ a technique that divides the decision-making process into a first-stage master problem wherein one set of decisions are determined, after which a set of separable second-stage problems are based on the first-stage decisions. The master problem determines an optimistic assessment of the costs incurred in the second stage; this assessment is then refined by a set of linear functions called cutting planes, which provide cause-and-effect relationships between the first- and second-stage decisions. However, these cutting planes are often ineffective in practical problems.

The improvement of these cutting planes is the subject of this proposal. The investigator encountered a surprising feature of some bi-level problems, in which an exponential number of cutting planes may need to be generated at each stage. By reformulating the first-stage problem using slightly more variables and constraints, it was possible to capture *all* of these cutting planes with a *single* inequality in the new variable space. Problems that were thought to be impossible to optimality are shown to in fact be solvable within reasonable computational limits.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE		11	19b. TELEPHONE NUMBER *(include area code)*

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39.18

Final Report: Enhanced Cutting Plane Techniques for Bi-Level Optimization Algorithms
Grant Number: FA9550-07-1-0404

PI: J. Cole Smith
University of Florida
Department of Industrial and Systems Engineering
PO Box 116595
Gainesville, FL 32607

Abstract (non-technical):

Bi-level optimization algorithms have been developed for use in a variety of problems, including network design and defense systems and optimization under uncertainty, and network interdiction applications. These algorithms often employ a technique that divides the decision-making process into a first-stage master problem wherein one set of decisions are determined, after which a set of separable second-stage problems are based on the first-stage decisions. The master problem determines an optimistic assessment of the costs incurred in the second stage; this assessment is then refined by a set of linear functions called cutting planes, which provide cause-and-effect relationships between the first- and second-stage decisions. However, these cutting planes are often ineffective in practical problems.

The improvement of these cutting planes is the subject of this proposal. The investigator encountered a surprising feature of some bi-level problems, in which an exponential number of cutting planes may need to be generated at each stage. By reformulating the first-stage problem using slightly more variables and constraints, it was possible to capture *all* of these cutting planes with a *single* inequality in the new variable space. Problems that were thought to be impossible to optimality are shown to in fact be solvable within reasonable computational limits.

Overview

The initial purpose of this brief seven-month proposal was to launch an investigation into methods by which the efficacy of bi-level optimization algorithms can be improved by examining a new technique that we have developed called "exponential inequality capture."

Bi-level optimization algorithms have been developed for use in a variety of problems, including network design, scenario-based stochastic programming, and network interdiction applications. These algorithms often employ a technique that divides the decision-making process into a first-stage master problem wherein "linking" variable values are determined, after which a set of separable second-stage subproblems are solved. The master problem determines an optimistic assessment of the second-stage objective; this assessment is then refined by a set of cutting planes, which provide "cause-and-effect" relationships between the linking decisions and second-stage objective function. However, these cutting planes are often too weak to be used for practical problems in which binary variables appear in the first stage.

The tightening of these cutting planes by an innovative *master problem reformulation* process was the subject of this proposal. The investigator has encountered a surprising feature of some bi-level problems, in which an exponential number of nondominated cutting planes (using Benders decomposition) can be generated at each stage. However, by (a) reformulating the master problem in a (polynomially sized) higher-dimensional space, and (b) linking the master problem and subproblems via these new variables, it is possible to imply *all* of the exponentially many cutting planes in the original variable space with a *single* inequality in the new variable space. This result implies that previous problems that were thought to be impossible to solve via exact optimization may in fact be solvable within reasonable computational limits.

We believe that the completion of this exploratory research has resulted in foundations that have clear benefit to the Air Force. We will give one such example in the following section, which is written in the context of product introduction but has a very relevant interpretation as an asset acquisition problem under competition. We will also demonstrate that this technique has clear application to multi-level optimization, game theory, and nonconvex optimization.

Illustration of Exponential Inequality Capture

To begin this report, we consider a two-person game played by two companies who can potentially introduce products from the same set of product designs, N, into a market. One firm, the *leader*, first introduces a set of products that belong to N. Then the other firm, the *follower*, produces its line of products that also belong to N. The market is assumed to be exclusively divided into a set of segments, M. Segment $i \in M$ has an (ordered) product preference list, $O_i = (p_i^1, \ldots, p_i^{k(i)})$, where $p_i^j \in N$, $\forall j = 1, \ldots, k(i)$, are arranged such that segment i will purchase the first product that is available in the list. That is, if products $p_i^1, \ldots, p_i^k$, $k < k(i)$, have not been introduced to the market, while product p_i^{k+1} has been, then consumers in market segment i will select product p_i^{k+1}. Furthermore, if no products in O_i have been introduced in the market, then segment i will not purchase any product.

For each $j \in N$, the leader and follower incur fixed costs b_j and c_j, respectively, for introducing products to the market. We assume that the leader has a budget of B, and the follower a budget of C, in the total expenditure for product introduction. Furthermore, assume that the revenue, r_{ij}, yielded from each market segment $i \in M$ consuming product $j \in N$ is known *a priori*. If both the leader and the follower offer product j, then we assume that $100\rho_{ij}\%$ of the demand for product j by market segment i is supplied by the leader's product, while the follower satisfies the rest. For convenience, let H_{ij} represent the set of markets that customer j prefers more than market i.

The objective of this particular problem is to maximize the leader's profit under the worst-case scenario in which the follower acts as a predator, i.e., attempts to minimizes the leader's revenue. In a military setting, this may correspond to a scenario in which one military unit acts first to secure certain bases or strategic areas. The follower may then act next to reply to the leader's actions, either by securing alternative areas, or challenging the leader for areas they initially established. (The success of competing for a secured area would be given by the ρ_{ij} parameters.) The "customers" here would be access to assets on a battlefield, which would go to the military unit that was nearest and would thus control the area.

Let x_j and y_j, for all $j \in N$, denote binary variables of the leader and the follower, respectively. If $x_j = 1$ ($y_j = 1$), then the leader (follower) introduces product j. Else, product j is not offered by the respective players. Furthermore, let w_{ij} be a variable that denotes the leader's actual revenue obtained from market segment $i \in M$ that buys product $j \in O_i$ after the follower's penetration.

Note that given values of $\bar{x}$ and $\bar{y}$ for x and y, respectively, the revenue collected by the leader is given by the optimal solution to the following problem.

$$
\begin{array}{llll}
\text{Maximize} & \displaystyle\sum_{i \in M} \sum_{j \in O_i} w_{ij} & & \text{(1a)} \\
\text{Subject to:} & w_{ij} \le r_{ij}(1 - \bar{x}_k) & \forall\, i \in M, j \in O_i, k \in H_{ij} & \text{(1b)} \\
& w_{ij} \le r_{ij}(1 - \bar{y}_k) & \forall\, i \in M, j \in O_i, k \in H_{ij} & \text{(1c)} \\
& w_{ij} \le r_{ij} \bar{x}_j & \forall\, i \in M, j \in O_i & \text{(1d)} \\
& w_{ij} \le r_{ij}(1 - (1 - \rho_{ij}) \bar{y}_j) & \forall\, i \in M, j \in O_i & \text{(1e)}
\end{array}
$$

We can now formulate the product introduction problem as a three-stage model:

$$
\begin{array}{llll}
\text{Max}_{x \in X}\ -b^{\mathrm{T}}x + \min_{y \in Y}\ \max & \displaystyle\sum_{i \in M} \sum_{j \in O_i} w_{ij} & & \text{(2a)} \\
\text{s.t.:} & w_{ij} \le r_{ij}(1 - \bar{x}_k) & \forall\, i \in M, j \in O_i, k \in H_{ij} & \text{(2b)} \\
& w_{ij} \le r_{ij}(1 - \bar{y}_k) & \forall\, i \in M, j \in O_i, k \in H_{ij} & \text{(2c)} \\
& w_{ij} \le r_{ij} \bar{x}_j & \forall\, i \in M, j \in O_i & \text{(2d)} \\
& w_{ij} \le r_{ij}(1 - (1 - \rho_{ij}) \bar{y}_j) & \forall\, i \in M, j \in O_i & \text{(2e)}
\end{array}
$$

where $Y = \{y: \Sigma_{j \in N} c_j y_j \leq C, y_j \text{ binary}, \forall j \in N\}$. Letting α_{ijk}, β_{ijk}, λ_{ij}, and μ_{ij} denote dual variables associated with constraints (1b), (1c), (1d), and (1e), respectively, the dual to (1) is given as:

$$\text{Minimize} \quad \sum_{i \in M} \sum_{j \in O_i} \sum_{k \in H_{ij}} \left[r_{ij}(1-\bar{x}_k)\alpha_{ijk} + r_{ij}(1-\bar{y}_k)\beta_{ijk} \right] + \sum_{i \in M} \sum_{j \in O_i} r_{ij}\bar{x}_j \lambda_{ij} + r_{ij}(1-(1-\rho_{ij})\bar{y}_j)\mu_{ij} \tag{3a}$$

$$\text{Subject to:} \quad \sum_{k \in H_{ij}} (\alpha_{ijk} + \beta_{ijk}) + \lambda_{ij} + \mu_{ij} = 1 \qquad \forall\, i \in M, j \in O_i \tag{3b}$$

$$\alpha, \beta, \lambda, \mu \geq 0. \tag{3c}$$

Using this dual, the problem in (2) can be rewritten as follows.

$$\text{Max}_{x \in X} \; -b^{\mathrm{T}}x + \min \sum_{i \in M} \sum_{j \in O_i} \sum_{k \in H_{ij}} \left[r_{ij}(1-x_k)\alpha_{ijk} + r_{ij}(1-y_k)\beta_{ijk} \right] + \sum_{i \in M} \sum_{j \in O_i} \left[r_{ij}x_j \lambda_{ij} + r_{ij}(1-(1-\rho_{ij})y_j)\mu_{ij} \right] \tag{4a}$$

$$\text{s.t.: Constraints (3b)–(3c)} \tag{4b}$$

$$y \in Y. \tag{4c}$$

Given x, the inner minimization problem in (4) is a bilinear integer program in which bilinear terms $y_k\beta_{ijk}$, and $y_j\mu_{ij}$ consist of one binary variable and one continuous variable. Furthermore, note that variables α, β, λ, and μ are constrained by (3b)–(3c), while the y-variables are (disjointly) constrained by the set Y. Let Π denote the set of extreme points to the polyhedron defined by (4b)–(4c). Then, we can rewrite (4) as:

$$\text{Max}_{x \in X} \; -b^{\mathrm{T}}x + \min_{y \in Y, (\alpha, \beta, \gamma, \mu) \in \Pi} \sum_{i \in M} \sum_{j \in O_i} \sum_{k \in H_{ij}} \left[r_{ij}(1-x_k)\alpha_{ijk} + r_{ij}(1-y_k)\beta_{ijk} \right] + \sum_{i \in M} \sum_{j \in O_i} \left[r_{ij}x_j \lambda_{ij} + r_{ij}(1-(1-\rho_{ij})y_j)\mu_{ij} \right]. \tag{5}$$

Since the objective function in the inner minimization problem of (5) is affine in terms of x, we can employ a cutting plane algorithm in the spirit of Benders decomposition. The Relaxed Master Problem (RMP) takes on the form

$$\text{Maximize} \quad -b^{\mathrm{T}}x + z \tag{6a}$$

$$\text{Subject to:} \quad z \leq \sum_{i \in M} \sum_{j \in O_i} \sum_{k \in H_{ij}} \left[r_{ij}(1-x_k)\bar{\alpha}_{ijk} + r_{ij}(1-\bar{y}_k)\bar{\beta}_{ijk} \right] + \sum_{i \in M} \sum_{j \in O_i} \left[r_{ij}x_j \bar{\lambda}_{ij} + r_{ij}(1-(1-\rho_{ij})\bar{y}_j)\bar{\mu}_{ij} \right] \quad \forall \; (\bar{y}, \bar{\alpha}, \bar{\beta}, \bar{\lambda}, \bar{\mu}) \in Y \times \Pi \tag{6b}$$

$$x \in X, \tag{6c}$$

while the subproblem is given by the inner minimization problem in (4).

The problem with the above formulation is that the cutting planes produced from this algorithm turn out to be very weak. The method by which we strengthen those inequalities embodies the results from our proposed research.

Note that the cutting planes (6b) can be interpreted as the statement of an upper bound on z, which can be modified by the selection or deselection of certain x-variables associated with positive α- and λ-values. The selection of these dual variables thus guides the decomposition search procedure by encouraging the selection of x-variables that maximize z.

Hence, setting $\alpha_{ijk} = 1$ for some $i \in M, j \in O_i$, and $k \in H_{ij}$ in the solution of (3) yields a term that implies that setting $x_k = 0$ can increase the right-hand-side by r_{ij}. However, if there were at least two terms $k_1, k_2 \in H_{ij}$ such that $\bar{x}_{k_1} = \bar{x}_{k_2} = 0$ in the previous solution, then setting either $\alpha_{ijk_1} = 1$ or $\alpha_{ijk_2} = 1$ does not accurately convey the condition that *both* x_{k_1} and x_{k_2} must equal zero in order to possibly capture segment i by introducing product j.

The existence of alternative optimal dual solutions in this case implies that several cuts can be generated from each $\bar{x}$-solution passed to the subproblem. *Worse, we find that there may exist an exponential number of these cuts, and that these cuts do not dominate one another.* This behavior creates a vital need for the research initiated in this proposal.

Given the problems explained above, we developed a method for removing the difficulties stemming from the presence of alternative optimal dual solutions. Consider the following reformulation of the master problem. For each segment/product pair $i \in M$, $j \in O_i$, define variables u_{ij} to equal one if j is the highest-preference product for market segment i that has been introduced by the leader, and zero otherwise. That is, u_{ij} equals to $x_j \prod_{k \in H_{ij}} (1 - x_k)$.

We linearize each of these terms using the following constraints:

$$u_{ij} \leq x_j \tag{7a}$$
$$u_{ij} \leq (1 - x_k) \quad \forall\, k \in H_{ij} \tag{7b}$$
$$u_{ij} \geq x_j + \sum_{k \in H_{ij}} (1 - x_k) - |H_{ij}| \tag{7c}$$
$$u_{ij} \geq 0. \tag{7d}$$

Problem (1) then becomes

$$\text{Maximize} \quad \sum_{i \in M} \sum_{j \in O_i} w_{ij} \tag{8a}$$
$$\text{Subject to:} \quad w_{ij} \leq r_{ij} \bar{u}_{ij} \qquad \forall\, i \in M, j \in O_i \tag{8b}$$
$$w_{ij} \leq r_{ij}(1 - \bar{y}_k) \qquad \forall\, i \in M, j \in O_i, k \in H_{ij} \tag{8c}$$
$$w_{ij} \leq r_{ij}\left(1 - (1 - \rho_{ij})\bar{y}_j\right) \quad \forall\, i \in M, j \in O_i, \tag{8d}$$

and taking its dual (with α_{ij} now associated with (8b), β_{ijk} associated with (8c), and μ_{ij} associated with (8d)) now yields:

$$\text{Minimize} \quad \sum_{i \in M} \sum_{j \in O_i} r_{ij} \bar{u}_{ij} \alpha_{ij} + \left[\sum_{k \in H_{ij}} r_{ij} (1 - \bar{y}_k) \beta_{ijk} \right] + r_{ij} \left(1 - \left(1 - \rho_{ij}\right) \bar{y}_j\right) \mu_{ij} \tag{9a}$$

$$\text{Subject to:} \quad \alpha_{ij} + \sum_{k \in H_{ij}} \beta_{ijk} + \mu_{ij} = 1 \qquad \forall\, i \in M, j \in O_i \tag{9b}$$

$$\alpha, \beta, \mu \geq 0. \tag{9c}$$

The resulting cutting plane is given by

$$z \leq \sum_{i \in M} \sum_{j \in O_i} r_{ij} u_{ij} \bar{\alpha}_{ij} + \left[\sum_{k \in H_{ij}} r_{ij} (1 - \bar{y}_k) \bar{\beta}_{ijk} \right] + r_{ij} \left(1 - \left(1 - \rho_{ij}\right) \bar{y}_j\right) \bar{\mu}_{ij} \,. \tag{10}$$

Smith et al. [3] provide details on deriving optimal duals to this problem in linear time, and demonstrate that an exponential number of cutting planes of the form (6b) can be implied by a single inequality of the form (10). This strength is due to the addition of only a quadratic number of new variables in the master problem. Essentially, the exponential cut capture is due to the fact that we can replace each u-variable with one of the x-variables for which it is a surrogate, and there can be an exponential number of different replacement choices, each leading to a distinct inequality in the x-space.

In comparing the x-variable approach with the u-variable approach, we derived 11 sets of test problems having different values of $|N|$, $|M|$, and $|O|$, along with different parameter settings and ratios (see [3] for a full description of these instances). A brief summary of the most effective method based on cuts (6b) (the "x-cuts") versus the most effective method based on (10) (the "u-cuts"), is given in Table 1, and conclusively demonstrates that master problem reformulation is extremely effective in this application.

Table 1: Average CPU time for eleven different problem profiles, where each entry is an average over ten problem instances

Problem Profile	x-cuts	u-cuts
1	122.9	0.4
2	569.9	1.1
3	1420.2	6.6
4	21.0	1.0
5	1579.0	2.7
6	1300.5	5.3
7	213.9	0.6
8	1445.0	1.1
9	46.9	1.7
10	295.4	1.8
11	8.7	1.2
Average	638.5	2.1

Additional Progress

The remainder of this report regards progress made on other related projects dealing with bi-level optimization via variants of mathematical programming decomposition algorithms.

Convexification strategies in solving risk management problems. In [1], we consider a class of two-stage stochastic risk management problems, which may be stated as follows. A decision-maker determines a set of binary first-stage decisions, after which a random event from a finite set of possible outcomes is realized. Depending on the realization of this outcome, a set of continuous second-stage decisions must then be made that attempt to minimize some risk function.

This risk threshold can be very simple: In one variant, there exists a single specified risk threshold value, which when exceeded under any recourse scenario, results in an additional objective penalty being incurred. The overall objective function thus depends on the cost of the first-stage decisions, plus the expected second-stage risk penalties. For instance, we can use this modeling technique to model the situation in which we have a limited budget that cannot be exceeded in the future without significant penalty. We would need to have a binary variable in each scenario's problem to indicate whether or not this penalty is incurred, thus preventing the use of Benders decomposition as described above.

In another variant, we consider a hierarchy of multiple risk levels along with associated penalties for each possible scenario. This may correspond to the case of gradually increasing penalties for missing certain target values. In construction situations, penalties increase in a nonlinear fashion due to lateness, which can be handled by our modeling scheme (since lateness is a discrete value). Also, these penalties can refer to varying costs of recourse schemes, e.g., to make up a shortage in supply for logistics scenarios.

In each case, we develop a mixed-integer 0-1 programming model and adopt an automatic convexification procedure using the Reformulation-Linearization Technique (RLT) to recast the problem into a form that is amenable to applying Benders' partitioning approach. As a principal computational expedient, we show in [1] how the reformulated higher-dimensional Benders' subproblems can be efficiently solved via certain reduced-sized linear programs in the original variable space.

Ad-hoc Cutting Planes for Integer and Nonlinear Subproblems. We contribute to the growing interest in creating multiple levels for optimization problems in order to improve their solvability. Some problems are very difficult because of scale and/or presence of nonlinear terms or integer restrictions in the problem. Occasionally, these difficulties can be effectively mitigated by splitting the problem into two phases, in order to permit more efficient handling of the complicating problem aspects.

In [2], the PI considers a problem in tournament scheduling. In particular, the problem regards the assignment of collegiate baseball teams to positions within a type of "knockout" tournament, where a single winner advances from a set of mini-tournaments until the remaining teams reach a neutral site location for a final championship. The tournament has three rounds, with optimization occurring to minimize teams' travel in rounds 1 and 2 (since the 3^{rd} round would be a fixed location). The tournament travel in round 1 is deterministic. However, the tournament travel in round 2 is a nonlinear function of round 1 team placement decisions, and is subject to stochasticity. While this problem could be modeled as a single nonlinear integer programming formulation, it has a very weak relaxation and cannot practically be solved to optimality (especially within the time limits that would be required to use it in a real setting). Instead, we split the problem into a two-stage problem. The second-stage problem becomes a linear integer program that can be quickly solved, with the nonlinearities being resolved due to the splitting of the problem. The PI [2] gives a combinatorial cutting-plane technique that is shown to quickly achieve a small optimality gap and produce tournament structures that are arguably superior to the ones derived over the last few years.

The primary difficulty in passing inequalities in [2], and in other problems involving binary variables in subproblems, is that duality information is not readily available from the second stage, and so the traditional Benders inequalities are necessary, but not sufficient, to force the algorithm to converge to an optimal solution. In response to this, a set of very weak (but valid) cutting planes was proposed roughly 15 years ago by Laporte and Louveaux. In [6], the PI and his student investigate a problem in facility location and activation, which is modeled as a two-stage problem, as described below.

A traditional facility location/allocation problem could be described as follows. In a first stage, a set of facilities must be located at some discrete points in a network, and each location incurs some fixed charge. Demands are treated as uncertain, but after realizing certain demands on the network, flows are then shipped across the network according to given costs for shipping along network arcs, and subject to capacities on these arcs. (In fact, the costs and capacities, and therefore the very existence of these arcs, may also be uncertain.) This setup corresponds well to the two-stage stochastic programming setup, because the flow problems are all solvable by linear programming, which yields duality information and permits the generation of Benders inequalities.

However, the classical problem ignores cases in which facilities are located with one setup charge, but another fixed charge must be paid in order to utilize a located facility. The optimization challenge posed in this [6] is inherent in many military planning situations. A long-range strategic plan may initially capture certain vital areas in some theater, which could then be pursued as forward bases for future action in the campaign. The first-stage "facility location" here would represent the vital areas to be secured early in the conflict. The second-stage decisions would represent the "activation"' of the facilities (by mobilizing troops and establishing advanced control centers), plus the costs associated with deploying troops from these established bases. (Note that advanced bases cannot be established in vital areas that were not previously secured.)

As another example, consider the location of hurricane relief shelters in hurricane-prone regions. If a hurricane were to strike a certain region, some shelters in the region would be opened and staffed, incurring a fixed activation cost. In the recourse problem, once a hurricane has struck a region, we are limited to the facilities that have been located in the first stage to address the needs of the hurricane victims. The variable flow costs would be proportional to the number of people the shelter serves in a given scenario, which could be modelled by sending flows from the shelter to those communities seeking aid.

The cutting planes prescribed in [6] are among the first to examine a particular problem having integer second-stage variables and prescribe valid inequalities that are at least as tight as the standard Laporte-Louveaux inequalities, due to the fact that they have been tailored to the particular application under study. (The technique employed is to essentially reverse-engineer flow solutions to determine the maximum benefit that could be obtained from locating a new facility in the first stage.)

Embedding Bi-Level Cuts in Constraint Programming. While the use of integer programming for solving difficult combinatorial optimization problems is well-established, there are several constructs that are difficult to express using linear constraints and integer variables. In some circumstances, solving a problem by constraint programming is a much more elegant algorithm that uses less computational (memory and CPU time) resources than integer programming. (Essentially, constraint programming is an implicit enumeration algorithm that uses advanced fathoming rules to limit its search space without risking the possibility of failing to identify an optimal solution.)

Rarely is the choice between these two methodologies clear, and traditionally, the problems are solved by whichever technique is preferred by the researcher. However, there are several structures that are particularly amenable to solution by mathematical programming, and many others that are better suited for constraint programming. Our contributions have sought to split notoriously difficult problems into two stages, even if the problem is more naturally modeled as a single problem. The advantage that we seek is to split the problem into one structure that is easily solvable by integer programming, and the other into a structure that is easily solvable by constraint programming, using bi-level inequalities to pass information between the two.

One key application of these strategies arises in [4] and [7]. Over the past decade, Intensity Modulated Radiation Therapy (IMRT) has developed into the most successful external-beam radiation therapy delivery technique for many forms of cancer. This is due to its ability to deliver highly complex dose distributions to cancer patients that enable the eradication of cancerous cells while limiting damage to nearby healthy organs and tissues. Patients treated with IMRT therefore often experience a higher chance of cure and suffer from fewer side effects of the treatment.

External-beam radiation therapy is delivered from multiple angles by a device that can rotate around a patient. The use of multiple (typically 3 – 9) angles is one of the tools that allow for the treatment of deep-seated tumors while limiting the radiation dose to surrounding functioning organs. IMRT is a more powerful therapy that modulates beam intensity, and dynamically shapes beams with the help of a multileaf collimator (MLC) system. Such systems can dynamically form many complex apertures by independently moving leaf pairs that block part of the radiation beam.

The optimization at problem becomes one of arranging the leaf pairs in order to deliver the right total dosage to a patient, as given by an input matrix called a *fluence map*. While some very small instances (e.g., roughly 5 x 5) have been solved to optimality using integer programming, no clinical-sized instance (e.g., roughly 20 x 20) had been solved, even when allocating may hours of computational time.

Our approach was to determine a set of radiation intensities to deliver in an integer programming phase, followed by a constraint programming phase that attempts to match the given intensity values to leaf-pair positions that it generates. If not successful, the constraint programming phase passes back a cutting plane to the integer programming phase, ensuring the eventual convergence of the approach. This unique approach was surprisingly able to solve 23 out of 25 instances to optimality within 30 minutes of computational time.

A similar approach was stated in [5] for a problem in telecommunication network design. In this problem, we actually formulate a three-stage problem in order to shift an integer programming network design phase to the first two stages (solved by a combination of integer and constraint programming), and then consider the impact of this design on a set of future scenarios in a third-stage integer programming phase, from which a set of combinatorial Benders inequalities are generated. This approach again dominates all other "pure algorithm" strategies as demonstrated by the computational results.

Future Research

This initial study, while busy, is still the proverbial tip-of-the-iceberg in this field of study. We are formalizing the situations in which exponential cut capture is possible, and when it is useful. We intend to continue to investigate the success of broader cutting-plane ideas, and how they can be used within multiple-algorithm settings such as the hybrid integer programming/constraint programming ideas presented above. Future funding has been secured to continue this project, and as a result, we will continue to provide results on bi-level optimization theory and methodology.

References:

The following papers are mentioned in the final report, and were partially or fully supported by this funding. The progress of these papers' publication can be seen at http://www.ise.ufl.edu/cole, and they are available either from the publisher, on-line at the given address, or from the investigator by contacting him at cole@ise.ufl.edu.

[1] Sherali, H.D. and Smith, J.C., "Two-Stage Stochastic Risk Threshold and Hierarchical Multiple Risk Problems: Models and Algorithms," to appear in *Mathematical Programming, Series A*.

[2] Smith, J.C., "Organization of a College Baseball Tournament," to appear in *IMA Journal of Management Mathematics*.

[3] Smith, J.C., Lim, C., and Alptekinoglu, A., "Optimal Mixed-Integer Programming and Heuristic Methods for a Bilevel Stackelberg Product Introduction Game," submitted to *Naval Research Logistics*.

[4] Taskın, Z.C., Smith, J.C., Romeijn, H.E., and Dempsey, J.F., "Optimal Multileaf Collimator Leaf Sequencing in IMRT Treatment Planning," submitted to *Operations Research*.

[5] Taskın, Z.C., Smith, J.C., Ahmed, S., and Schaefer, A.J., "Cutting Plane Algorithms for Solving a Robust Edge-Partition Problem," submitted to *INFORMS Journal on Computing*.

[6] Penuel, J. and Smith, J.C., "An Integer Decomposition Algorithm for Solving a Two-Stage Facility Location Problem with Second-Stage Activation Costs," submitted to *Naval Research Logistics*.

[7] Taskın, Z.C., Smith, J.C., and Romeijn, H.E. "Mixed-Integer Programming Techniques for Decomposing IMRT Fluence Maps Using Rectangular Apertures," submitted to *Annals of Operations Research*.